Summary of wrap up discussion

Fowler, Floyd Jackson Jr.

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

SUMMARY OF WRAP UP DISCUSSION

FLOYD JACKSON FOWLER, JR.

The final session of the workshop was devoted to two topics: what is the state of our current knowledge about how to evaluate questions and what are the priorities for needed research? The following is a brief summary of the conclusions from that session.

State of current knowledge

Although there is still much that we need to learn, progress has been made since our last meeting in documenting the value of question testing and in producing empirically-based generalizations about how to conduct testing.

1. The study conducted by Forsyth, Rothgeb, and Willis (2004) clearly demonstrated that pretesting does identify real question problems. Moreover, Fowler (2004) reports data that strongly support the position that problems identified in cognitive testing can have important effects on the data, the quantitative estimates made from surveys.

2. Expert reviews, cognitive testing, and field pretests with behavior coding all have contributions to make to the evaluation of questions. No one method is adequate to identify all of the various kinds of problems that questions can have. The fact that each method has the ability to identify some kinds of problems and not others should not be considered "flaws" in the method. Rather, we need to recognize that the approaches are complementary and that multiple methods should be used to comprehensively evaluate questions.

3. Evaluating the usability of survey instruments is as important as evaluating the question wording. The best techniques for evaluating usability and question wording are different, again leading to the conclusion that multiple testing approaches are needed.

4. With respect to the particular techniques for doing cognitive testing, DeMaio and Landreth's research (2004) supports the following conclusions:

    a. Listening to tapes of cognitive interviews to help prepare reports of the results increases the number of issues identified and almost certainly increases the value of the cognitive testing.

b. Cognitive interviews are done by researchers with Ph.D.s, specialists in question design, and regular interviewers with special training. There is no evidence that interviewers with any one of these backgrounds is consistently superior to the others.

c. Cognitive interviews can rely on think-aloud techniques and probes that are asked after questions have been answered. Sometimes probes are scripted; at other times the interviewers create their own probes or use a combination of scripted and ad hoc probes. There is no evidence that any of these variations is consistently superior to the others.

Needed Research

The list of things that are known about question evaluation is shorter than the list of areas in which further research is needed. There are many aspects of how to test questions about which we need to learn more:

1. There is much to learn about how to use each of the methods of question evaluation that were discussed:

   a. We need descriptive studies of how organizations are currently doing the various kinds of testing. We know there is variation, but better documentation of the variation that currently exists would help to define a research agenda for how best to study evaluation techniques.

   b. We also need to do more studies documenting the strengths and limitations of the various approaches to evaluating questions and survey instruments.

   c. For cognitive testing, we need critical evaluation of the various kinds of probes that are used, to learn which are the best and most productive

   d. We also need more research on how best to use the results from cognitive testing to improve questions and data quality: How interviewers should record and report what happens (audiotapes, videotapes, write ups of individual interviews, summaries across several interviews); who (i.e. people with what sort of credentials or role in the survey) are the best people to view or listen to the actual interviews; what is the best process for taking the results of the cognitive interviews and using them to revise question wording?

   e. Behavior coding has concentrated on four or five behaviors (question reading, probing, requests for clarification, interruptions, and inadequate answers) to provide evidence of question problems. Research to identify other significant behaviors and to document the relationships between behaviors and undesirable question features will add to the value of behavior coding.

   f. Expert reviewers need better empirically-based generalizations about how (and how much) observable question characteristics affect usability and data quality.

2. There is good reason to think that testing techniques may need to be adapted to the requirements of certain kinds of surveys and populations. For example:
   a. Think-aloud interviews clearly seem to work better with some respondents than others
   b. There is anecdotal reason to think that the best cognitive probes may vary from respondent to respondent. We need studies to document and understand those relationships
   c. The best approaches to testing questions for surveys of individuals may differ from the best approaches to testing questions and instruments for establishment surveys.
   d. We have much to learn about how mode of administration affects the best way to evaluate questions and instruments. This is not just a matter of making sure that usability is evaluated. It is specifically addressing the fact that the way a question is understood and answered may vary by whether the respondent hears it on the telephone, reads it from a written page or reads it on a computer screen. Testing methods must be adapted to detect the problems that are specific to the mode of administration. At the moment, we do not have generalizations about how to adapt our testing to the mode of administration.
   e. Usability testing has emerged as one of the most underdeveloped aspects of survey instrument evaluation. Its importance has risen because computer-based data collection is on the rise. It is understood that respondent willingness to complete surveys that are self administered, on paper or on a computer, is affected by how easy it is to do. More studies are needed about how to efficiently and effectively identify problems that real respondents will have when they try to complete a survey.
3. To do these studies, we need to develop some new tools.
   a. One of the major barriers to studying question evaluation strategies is the absence of good measures of success. The studies to date have largely counted and compared the number of "problems" identified by testing, with no real way of knowing whether a “problem” actually had any effect on the survey estimates. Split-ballot tests are one approach to assessing the effects of alternative question wording on data, though knowing which question is “best” depends on having a strong theory or independent data if results are different. An even better design is to have measures in the survey instrument or external data to directly assess the validity of the answers to alternative questions. To date such validation studies have been rare, but we need them to evaluate the effectiveness of our testing and the questions changes that they produce.
   b. We need a better typology of question problems – one specifically geared to question evaluation techniques. Tourangeau’s (1984) sorting of issues into comprehension, retrieval, coding and answering was a start. However, we need a

more refined system that would enable researchers to match the kind of problem they are potentially concerned about with a particular approach to testing or probing in a cognitive interview.

c. We also need a typology of usability problems. Following instructions, seeing and understanding all the relevant parts of a question, knowing how to use computer aids, and simply knowing how to navigate through a set of questions are all part of usability. We need a parsimonious and useful list of usability issues, so we can be sure we are using testing procedures that will identify problems of the various types that matter.

d. The importance of usability needs to be assessed. It is a reasonable hypothesis that problems with usability adversely affect comprehension directly, as well as affecting willingness or ability to answer questions (leading to item nonresponse). However, those relationships have not been documented.

e. Finally, we need a better set of generalizations about features of questions that adversely affect measurement. Question testing identifies question or instrument features we think are problems. However, if we have stronger, evidence-based generalizations about what question features to avoid, we would have fewer bad questions to start with and the effectiveness of improving questions that were found wanting via testing would be greatly enhanced.

References

Forsyth, B., Rothgeb, J., and Willis, G., 2004: Does Pretesting Make a Difference? in S. Presser et al. (eds.) Questionnaire Development Evaluation and Testing Methods, New York: Wiley.

Fowler, F. J., 2004: Getting Beyond Pretesting and Cognitive Interviews: The Case for More Experimental Pilot Studies, in S. Presser et al. (eds.) Questionnaire Development Evaluation and Testing Methods, New York: Wiley.

DeMaio, T., and Landreth, A., 2004: Do Different Cognitive Interview Methods Produce Different Results? in S. Presser et al. (eds.) Questionnaire Development Evaluation and Testing Methods, New York: Wiley.

Tourangeau, R., 1984: Cognitive Science and Survey Methods. pp. 73-100 in: T. Jabine/ M. Straf/J. Tanner/R. Tourangeau (eds.) Cognitive Aspects of Survey Design: Building a Bridge Between Disciplines Washington: National Academy Press.

Contact

Floyd Jackson Fowler, Jr.
Center for Survey Research
University of Massachusetts, Boston
100 Morrissey Blvd.
Boston, MA 02125 -USA
email: floyd.fowler@umb.edu